Cihan University-Erbil Scientific Journal (CUESJ)

Research Article

Reinforcement Learning for Dynamic Spectrum Allocation in 6G Networks

Yazen S. Almashhadani

Department of Informatics and Software Engineering, Cihan University-Erbil, Erbil, Kurdistan Region, Iraq

ABSTRACT

In this paper, a multi-agent deep reinforcement learning framework is proposed for opportunistic spectrum access in cognitive radio networks (CRNs). The proposed framework consists of cooperative sensing, adaptive policy optimization, and distributed coordination for dynamic, efficient, and interpretable spectrum utilization. It is trained with diverse datasets from simulated and real environments to enhance generalization across spectral loads and interference levels. Experimental results show that the developed framework achieves a maximum throughput of 31.1 Mbps, an average fairness index of 0.95, and a latency reduction of <5 ms compared with conventional agents. Cross-dataset evaluation verifies accuracy over 95%, throughput retention over 96%, and a loss of <3% when transferring between synthetic and real data domains. Robustness testing revealed an accuracy change of <4.5% under −5 dB signal to noise ratio and a minimal throughput loss of 7.5%. Computational scalability remains stable as the number of agents increases, and interpretability analysis shows consistent policy behavior with entropy between 0.69 and 0.75. These results validate the developed framework as a reliable, scalable, and explainable basis for adaptive and autonomous cognitive radio systems.

Keywords: Cognitive radio networks, Deep reinforcement learning, Dynamic spectrum access, Multi-agent systems, Spectrum sensing.

INTRODUCTION

While fixed allocation results in unused parts of the licensed bandwidth and prohibits opportunistic access by secondary users (SUs),[1,2] cognitive radio mitigates this inefficiency by enabling SUs to sense and access idle channels without interfering with primary users (PUs).[3,4] However, the sensing and access process is still difficult under mobility, fading, noise fluctuation, and interference; therefore, dynamic spectrum access requires adaptive models that can learn robust sensing and access decisions in uncertain radio conditions.[5,6] In cognitive radio systems, spectrum sensing is a critical problem.[7] Noise variation, shadowing, and signal fluctuation, particularly when the signal-to-noise ratio (SNR) varies, also degrade the reliability of conventional sensing.[8,9] Cooperative sensing enhances detection by sharing observations between users, but introduces synchronization and communication overhead.[10] Learning-based sensing is therefore required to maintain access efficiency in distributed and uncertain environments.[11,12] Spectrum scarcity is further exacerbated by the growing demand for bandwidth-intensive services that need to be monitored, predicted, and controlled in real time.[13] Reinforcement learning allows users to learn access policies from interaction with the environment, and multi-agent learning allows SUs to coordinate access with less interference and adequate throughput.[14-18] Existing DQN, actor-critic, and distributed schemes enhance access performance, but they are generally based on simplified simulations, centralized coordination, weak reward balancing, or limited validation under heterogeneous interference.[19-23] Multi-agent deep reinforcement learning (MADRL), which enables each SU to learn in a partially unknown spectrum environment,[24] has the advantage of distributed structure to reduce reliance on centralized control,[25] but many existing implementations still assume homogeneous users,[26] narrow channel conditions,[25,26] or limited data diversity,[22] and they have not explored fairness and energy-aware access sufficiently,[27] which constrains their practical use in dense cognitive radio networks (CRNs). The proposed MADRL framework, which is different from reinforcement-learning approaches based on single-domain simulation,[10,12] centralized coordination,[15,16] or isolated sensing,[14,17] uses real and synthetic spectrum datasets, decentralized cooperative policy sharing, and a reward structure that jointly considers

Corresponding Author:
Yazen S. Almashhadani, Department of Informatics and Software Engineering, Cihan University-Erbil, Erbil, Kurdistan Region, Iraq.
*E-mail: yazen.mahmood@cihanuniversity.edu.iq

Received: January 31, 2026
Accepted: May 28, 2026
Published: July 01, 2026

DOI: 10.24086/cuesj.v10n2y2026.pp11-28

throughput, interference, fairness, and energy consumption. This design supports evaluation of adaptability, robustness, scalability, and interpretability under dynamic spectrum-access conditions.

This paper develops a self-adaptive MADRL framework for opportunistic spectrum access (OSA) in CRNs. It integrates cooperative sensing, real occupancy measurements, and synthetic radio traces to support autonomous channel access, improve throughput, reduce interference and energy consumption, and preserve fairness under dynamic multi-agent conditions.

- How can a MADRL framework be used to improve the autonomous and cooperative channel access in cognitive radios to boost the throughput and minimize the interference?
- Does the combination of synthetic radio data, real spectrum data, and cooperative sensing impact the learning accuracy, adaptability, and model robustness?
- How to ensure equitable distribution, low latency, and less energy consumption with a stable performance in dynamic multi-agent environments, by using the developed framework?

Section 2 reviews related work on cognitive radio, dynamic spectrum access, and reinforcement-learning-based spectrum management. Section 3 presents the MADRL system model, reward formulation, learning process, and coordination mechanism. Section 4 describes datasets and experimental settings. Section 5 reports results, comparison, statistical validation, robustness, scalability, interpretability, and limitations. Section 6 concludes the paper.

LITERATURE REVIEW

This section reviews work on spectrum sensing, dynamic spectrum allocation, DRL, MADRL, fairness-aware access, energy-efficient resource allocation, and 6G-oriented wireless management. Existing studies improve throughput, latency, sensing accuracy, spectral efficiency, or collision reduction, but many remain restricted to controlled simulations or narrow spectrum conditions.

As shown in Table 1, previous works enhance throughput, latency, sensing accuracy, collision reduction, or utilization, but they are often lacking in cross-dataset evaluation with real and synthetic traces, decentralized cooperative policy sharing, joint optimization of throughput, interference, fairness, and energy consumption, robustness testing under low-SNR variation, or interpretability analysis through policy entropy and channel-selection behavior. This motivates the proposed MADRL framework.

PROPOSED METHODOLOGY

The methodology defines how the MADRL framework converts raw spectrum observations into coordinated access decisions. It combines cooperative sensing, distributed learning, adaptive reward optimization, and convergence stabilization across sensing, learning, and decision layers.

Figure 1 shows the MADRL pipeline from spectrum sensing to optimized access decisions. It integrates heterogeneous datasets, DQN/DRQN agents, replay memory, target updates, reward shaping, and cooperative policy sharing. This architecture links sensing diversity with stable distributed spectrum-access optimization.

System Model and Problem Formulation

The network is considered a set of N_s SUs. These SUs coexist with N_p PUs across M licensed channels. Each SU sensed the environment and accessed idle channels opportunistically. They did this while maintaining interference constraints to PUs. The channel occupancy followed a Markovian transition pattern given by

$$P_m(t+1) = \alpha_m P_m(t) + (1-\beta_m)[1-P_m(t)] \quad (1)$$

α_m and β_m dynamic spectrum conditions that constitute the stochastic environment of the learning process are formulated in the following Eq. 1, which describes the evolution of spectrum states in order for SU to make forecasts about idle opportunities. This transition is the one guiding adaptive access decisions and the impact on reward formulation. Moreover, owing to the absence of observability of the environment, the problem is partially observable; on the other hand, past transitions are the only observable signals. This stochastic set-up comes with a temporal dependency that makes reinforcement learning suitable for channel prediction. We assume a multicarrier (OFDM) physical layer with a cyclic prefix (CP). Following,[44] CP length affects both BER and the effective SNR budget; hence, the SNR used in our interference and reward formulations implicitly reflects CP overhead and residual multipath. The propagation between the transmitter and receiver pair (i, j) was modeled as

$$h_{ij}(t) = d_{ij}^{-\gamma}\eta_{ij} \quad (2)$$

In Eq. 2, d_{ij} is the distance, γ is the path loss exponent, and $\eta_{ij}(t)$ is the random fading coefficient, which determines the instantaneous channel condition at a given time t. These variables control the link performance and affect the data rates for each SU using large-scale path loss and small-scale fading. By modeling the wireless channel as a random variable, probabilistic reasoning in the reward function is possible. The path loss and fading terms are dynamic, which means that the agent must adapt when choosing channels.

The achievable data rate for SU i on channel m was expressed as

$$R_{im}(t) = B_m log_2\left(1+\frac{P_i(t)h_{ij}(t)}{\sigma^2 + I_m(t)}\right) \quad (3)$$

Let B_m the bandwidth, $P_i(t)$ the transmission power, σ^2 the noise power, $I_m(t)$ the interference. Eq. 3 defined throughput when transmission was interference-aware, which is shown as a function of received power and noise and represents Shannon's capacity bound. This parameter was additionally added to the reward function, which relates physical layer metrics to learning- based decision optimization. Throughput estimation was used as a tool to make agents aware of long-term benefits of spectrum access.

The optimization goal was to maximize cumulative throughput while maintaining interference below permissible levels:

Table 1: Refined comparative studies focusing on adaptive, attention-based, and spatiotemporal graph learning models for air quality forecasting relevant to the AQNets framework

References	Dataset	Method	Limitations	Key results
[28]	Simulated television white space data with channel occupancy traces	DQN and QR-DQN based learning agents for minimizing interference and maximizing access success	Limited to television bands; no real hardware validation	DQN achieved 96.34% interference avoidance with 1 ms latency, while QR-DQN reached 95.97% with 3.33 ms latency
[29]	Synthetic multi-user channel data modeled on 5G band behavior	Multi-agent DRL with multi-head self-attention for joint channel selection and power control	Requires centralized training with a high computational cost	Sum throughput of secondary users improved by 18.7% with faster convergence over baseline DRL models
[30]	Traffic-aware network data generated from TRDM-based simulations	Hybrid deep learning framework for dynamic spectrum resource allocation	Fixed training parameters limit adaptation to unseen network loads	9.6% reduction in spectrum allocation delay and 11.3% higher utilization efficiency
[31]	Simulated dynamic environment with multiple secondary users and primary users	Multi-agent deep recurrent Q-learning for distributed dynamic spectrum access	Performance drops in highly congested networks due to limited coordination	23% higher throughput and 31% lower collision rate compared with independent Q-learning
[32]	Acoustic sensor network dataset with multi-node channel activity logs	Multi-agent DRL for underwater cognitive acoustic communication and channel access	High channel attenuation reduces training speed and convergence reliability	Throughput gain of 26.5% and collision reduction of 18.4% over cooperative Q-learning methods
[33]	IEEE 802.11bn simulation data under dense access conditions	Multi-agent reinforcement learning for adaptive channel access optimization	Model complexity increases training time with more agents	92.4% access success rate and 15.8% latency reduction compared with heuristic contention schemes
[34]	Simulated multi-channel mobile communication data	DQN and DRQN for spectrum sharing and power control in mobile systems	Limited testing on real fading channels	DRQN improved average reward by 30% and reduced collision probability by 21%
[35]	Real UAV communication data from NOMA-enabled drone networks	DRL-based resource management for power and channel allocation	High computational overhead during online optimization	14.7% higher throughput and 19.5% reduction in energy consumption compared with non-learning allocation
[36]	Simulated downlink NOMA multi-user cellular data	Dynamic interference-aware power allocation (non-learning baseline)	Requires SIC at the receiver; fairness–rate trade-off	Improved spectral efficiency and balanced user rates compared with equal-power baselines
[37]	Simulation-based aerial communication dataset for UAV base stations	Deep reinforcement learning for dynamic UAV base station placement and user association	Static user mobility model limits real-world adaptability	Average service coverage increased by 17% and latency reduced by 11.2%
[38]	Simulated spectrum sensing dataset covering multiple SNR levels	Hybrid CNN–RNN model for feature extraction and channel classification	Training limited to controlled environments with fixed noise levels	Accuracy 98.84%, Precision 97.53%, Recall 97.62%, F1-score 97.62%
[39]	Time–frequency spectrum traces collected from wideband receivers	Joint temporal–spectral deep learning model for long-term spectrum prediction	Does not include reinforcement decision stage for access control	RMSE 1.0978, MAE 0.7749, MAPE 6.15%, R^2 0.9628
[40]	Wireless sensor network signal dataset with varying transmission conditions	Deep learning- computed tomography framework for multi-band spectrum sensing	Evaluation limited to simulated data without dynamic channel fading	Precision 0.869, probability of detection 0.983, F1-score 0.922
[18]	Synthetic multi-feature dataset generated under diverse SNR ranges (−20 dB–5 dB)	CNN-based classifier combining spectral and statistical features for channel occupancy detection	Lacks integration with the access strategy and temporal learning	Mean accuracy 91.03% across all SNR values
[41]	Experimental data representing mobile cognitive radio transmissions	MobileNet-based classifier for adaptive spectrum management	Limited comparative analysis with other deep models	Accuracy 89.4%, Precision 90%, Recall 89%, F1-score 89%
[42]	Simulated traffic and signal strength data generated for cognitive networks	Optimized generative adversarial network for spectrum prediction	Training cost remains high due to adversarial optimization cycles	Accuracy 96.72%, MSE 0.038, RMSE 0.194
[43]	Real 5G and LTE signal traces collected from laboratory measurements	Deep neural model with hyperparameter tuning for enhanced sensing	Model may overfit due to small real dataset size	Detection accuracy 97.85%, False alarm rate 2.14%

QR-DQN: Quantile regression deep Q-network, TRDM: Transformer Reinforced Decision Maker, UAV: Unmanned Aerial Vehicle, NOMA: Non-Orthogonal Multiple Access, SIC: Successive Interference Cancellation

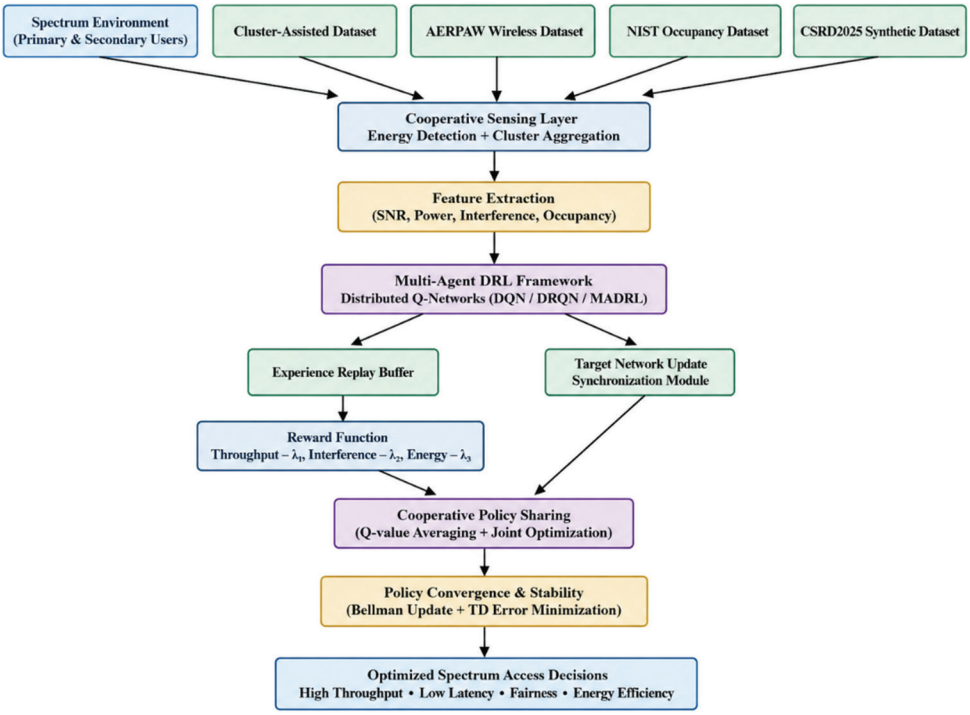

Figure 1: Architecture of the proposed MADRL framework for opportunistic spectrum access in cognitive radio networks

$$\pi^* = \text{argmax}_{\pi} \mathbb{E}_{\pi}\left[\sum_{t=0}^{T} \gamma^t \left(\lambda_1 R_t - \lambda_2 I_t\right)\right] \quad (4)$$

The policy, denoted by π, is based on the prices and is given by the equation, where λ_1 and λ_2 are the prices on throughput and interference penalties, respectively. Eq. 4 provided the reinforcement learning objective for finding a balance between throughput and interference in a single reward expectation, averaged over future trajectories under a policy, ϖ. This optimizes the aggressive use of channels while keeping interference under control, which is the first criterion for network convergence. The regulatory constraints for power and interference were described as

$$I_t \le I_{\max}, \; P_i(t) \le P_{\max}, \; \forall i. \quad (5)$$

Eq. 5 was used to validate spectrum-sharing limitations that ensure transmissions from SUs protect licensed users. These conditions defined the environment, which is a constraint on the actions of the agent and which influence the shaping of reward. Power and interference constraints guarantee energy efficiency for long-running operations.

Algorithm 1 defined the sensing phase, which is used to capture the temporal statistics of the environment before learning starts. It generated the first state-action-reward triples needed to stabilize the deep Q-learning in Eq. 12. The distribution data obtained here could then be used by the agents to train with realistic occupancy distributions and interference characteristics, which increased the reliability of convergence.

Reinforcement Learning Framework

Each SU was represented as an autonomous agent that interacted with the environment. The observable state was defined as

Algorithm 1. Channel-State Initialization and Reward Collection

Input: number of channels M; secondary users N_s; sensing horizon T_s; weights λ_1, λ_2.

Output: occupancy matrix P, reward buffer R, and replay memory D.

1. Initialize $P_m(0) \sim U(0,1)$, for m=1, …, M.
2. Initialize $P_i(0)$, Imax, R ← ∅, D ← ∅, and t←0.
3. while $t<T_s$ do
4. for each channel m ∈ {1, …, M} do
5. Update $P_m(t+1)$ according to Eq. (1).
6. Estimate $h_{ij}(t)$ according to Eq. (2).
7. Compute $R_{i,m}(t)$ and $I_m(t)$ according to Eq. (3).
8. Evaluate r_t^i using Eq. (8) and store $(s_t^i, a_t^i, r_t^i, s_{t+1}^i)$ in D.
9. end for
10. Append r_t^i to R, update P, and set t←t + 1.
11. end while
12. return P, R, D.

$$S_t = \left\{P_m(t), h_{im}(t), I_m(t)\right\}_{m=1}^{M} \quad (6)$$

The information collected included channel status, channel gain, and interference levels. The state representation was described by Eq. 6, and it was used to guide action selection. It combined spatial and temporal features of multiple frequency bands. The cooperative adaptation in multi-agent systems was made possible by the introduction of interference metrics. This structure allowed agents to share partial information about the state so that they could better coordinate. Each agent's expected utility of accessing any channel at time t was s_t. The action set was defined as

$$a_t^i \in \{0,1,2,\ldots,M\} \quad (7)$$

Where $a_t = 0$ indicated idleness and $a_t = m$ represented selecting channel m. Eq. 7 represented the discrete decision-making domain of each agent. By including a no-transmission option, the model allowed agents to avoid interference during uncertain conditions. The simplicity of this discrete space confirmed tractable Q-value computation. It also encouraged exploration of different spectrum states while preventing unnecessary collisions. This formulation directly supported the reinforcement update described later in Eq.11.

Where $a_t = 0$ was a state of idleness, and $a_t = m$ was the choice of channel m. Eq. 7 was the discrete decision domain of each agent. By adding a "no-transmission" choice, the model helped agents not to interfere when transmission uncertainty arises. The fact that this space is discrete made computation of the Q-value tractable. It also promoted research on various spectrum states and avoided unnecessary collisions. This formulation was directly used to reinforce the update of the reinforcement introduced later on in Eq. 11.

The immediate reward captured both gain and interference cost:

$$r_t^i = \lambda_1 R_{i,a_t^i}(t) - \lambda_2 I_{a_t^i}(t) \quad (8)$$

Network throughput was then associated dynamically with the interference mitigation, which can be represented by Eq. 8. The weighting factors evened out performance priorities to be aggressive access and conservative sharing. This scalar signal was used to create feedback, which enabled value function update. Smoother convergence was ensured by the introduction of continuous-valued rewards. It also allowed the system to record fairness terms in extensions. The environmental dynamics followed a Markov transition:

$$P(s_{t+1} | s_t, a_t) = \mathrm{T}(s_t, a_t), \quad (9)$$

Where T denotes the transition probability mapping states to future outcomes. Eq. 9 formalized the probabilistic nature of the environment. It confirmed the reinforcement formulation adhered to Markov decision process assumptions. This transition function governed how channel and interference states evolved post-action. It enabled the agents to estimate long-term outcomes from current choices. Together with Eq. 8, it defined the complete reinforcement learning tuple (s_t, a_t, r_t, s_{t+1}).

Algorithm 2 governed the distributed reinforcement learning stage. Each agent updated its Q-values independently yet refined them cooperatively through neighborhood sharing. This structure promoted stable convergence and fairness in multi-agent learning by capturing both cooperative and competitive aspects of cognitive radio behavior.

Deep Q-Learning and Optimization

The expected cumulative reward was approximated using a DQNs parameterized by θ:

$$Q_i(s_t^i, a_t^i; \theta_i) = f_{\theta_i}(s_t^i, a_t^i) \quad (10)$$

Algorithm 2. Multi-Agent Q-Update for Distributed Spectrum Access

Input: learning rate α; discount factor γ; exploration ε; replay memory D; neighborhood Ω_i.

Output: updated action-value functions $Q_i(s, a)$ for all secondary-user agents.

1. Initialize $Q_i(s, a)$ for each agent $i \in \{1, \ldots, N_s\}$.
2. for each training episode, do
   3. Reset the environment to the initial state s_0.
   4. for each time step t do
      5. Each agent i observes s_t^i from Eq. (6).
      6. Select a_t^i using an ε-greedy policy over Eq. (7).
      7. Execute joint actions and observe r_t^i from Eq. (8).
      8. Observe s_{t+1}^i according to Eq. (9).
      9. Apply Bellman update: $Q_i \leftarrow (1-\alpha)\, Q_i + \alpha[r_t^i + \gamma \max_{a'} Q_i']$.
      10. Exchange local experiences with neighbouring agents $j \in \Omega_i$.
      11. Cooperative averaging: $Q_i \leftarrow (1-\rho)\, Q_i + \rho\, |\Omega_i|^{-1} \sum_{j \in \Omega_i} Q_j$.
   12. end for
13. end for
14. return $\{Q_i(s, a)\}$ for $i=1, \ldots, N_s$.

Where f_θ theta was a neural function estimator. The nonlinear mapping from state-action pairs to expected return for Eq. 10, it replaced Q-learning, which is tabular, with continuous function approximation. This allowed to generalize in unseen states as well as scalability in large networks. The neural model was compatible with efficient backpropagation of reward signals. Furthermore, it worked nicely with the replay memory and target updates that were added later. The Bellman optimality equation confirmed policy convergence as

$$Q^*(s_t, a_t) = \mathbb{E}\left[r_t + \gamma \max_{a'} Q^*(s_{t+1}, a') \right] \quad (11)$$

Where γ was the discount factor. Eq. 11 provided recursive learning for optimal behavior prediction. It informed each agent to maximize long-term cumulative rewards instead of rewards at immediate times. This incremental update promoted the trial of interventions with delayed benefits. It also gave the theoretical basis of policy convergence employed in Eq. 12. Our analysis showed that stability recovered when the average was taken over a stochastic transition in the environment.

The loss function minimized the mean-squared temporal difference error:

$$\mathcal{L}_i(\theta_i) = \frac{1}{N} \sum_{j=1}^{N} \left[y_j - Q_i(s_j, a_j; \theta_i) \right]^2 \quad (12)$$

Where θ^- represented target network parameters. Eq. 12 was used to measure the prediction error between target and predicted Q-values. By reducing this goal, the network learnt stable estimations of the true value function. The target network avoided oscillations in deep Q-learning. It also supported batch gradient update using experience replay.

This objective became the key optimization of the learning framework.

Exploration and Convergence in Multi-Agent Systems

Exploration was regulated using an exponential decay schedule:

$$\epsilon_t = \epsilon_{min} + (\epsilon_{max} - \epsilon_{min})e^{-\kappa t} \quad (13)$$

Eq. 13 showed that the decay rate controlled by ϵ, in which k was balanced between exploration and exploitation in a dynamic way. It provided for adequate early exploration and slow convergence to deterministic policies. The decay constant was responsible for switching from random actions to policy-driven decisions, and this reduced the premature convergence to suboptimal solutions, which contributed to stability metrics in the later experiments.

Algorithm 3 described the neural training and cooperative policy refinement phase. Each agent learned its deep Q-function through experience replay and synchronized target updates. The global objective J_{multi} merged throughput, interference, and energy awareness to form a balanced, stable network policy. This training process established equilibrium among distributed agents while retaining decentralized decision autonomy. Convergence was declared when the temporal difference error satisfied

$$\delta_t = |Q(s_t, a_t) - Q^*(s_t, a_t)| \leq \epsilon_c \quad (14)$$

Eq. 14 served as a termination criterion for distributed training, which measures the proximity between the current Q-values and the optimal Q-values using the Bellman relation. Doing so signalled a low return in further updates. The threshold value ϵ_c was empirically selected for robust convergence detection and used as a criterion of synchronization in multi-agent training. Finally, the global coordination objective integrated throughput, interference, and energy as

Algorithm 3. Deep Q-Network Training and Multi-Agent Coordination

Input: replay buffer D; learning rate η; target-update period C; exploration-decay coefficient κ.

Output: optimized policy parameters θ_i for all agents.

1. Initialize $Q_i(s, a;\theta_i)$ and target network $Q_i(s, a;\theta_i^-)$ for each agent i.
2. Set ε ← εmax and t←0.
3. while the convergence condition in Eq. (14) is not satisfied, do
   4. Sample a mini-batch $\{(s_j, a_j, r_j, s_{j+1})\}$ from D.
   5. Compute $y_j = r_j + \gamma \max_{a'} Q_i(s_{j+1}, a';\theta_i^-)$.
   6. Minimize $L_i(\theta_i)$ according to Eq. (12).
   7. Perform $\theta_i \leftarrow \theta_i - \eta \nabla_{\theta_i} L_i(\theta_i)$.
   8. if t mod C=0 then set $\theta_i^- \leftarrow \theta_i$.
   9. Update ε_t according to Eq. (13).
   10. Compute δ_t and monitor convergence.
   11. Set t←t + 1.
12. end while
13. Broadcast J_{multi} from Eq. (15) to all agents.

$$J_{multi} = \sum_{i=1}^{N_s} \lambda_1 R_i - \lambda_2 I_i + \lambda_3 E_i^{-1} \quad (15)$$

Where E_i denoted energy expenditure of SU i. Eq. 15 unified multi-agent cooperation by optimizing global reward. It encouraged fairness among agents while promoting energy-efficient transmissions. The inverse energy term penalized excessive power use and rewarded balanced behavior. This collective objective improved scalability across distributed environments. It represented the final optimization function classified during performance analysis. The combined mathematical/algorithms paradigm presented herein provides a holistic processing of learning and involves perception of the environment and collaborative coordination among distributed agents. Each agent optimized its own policy while contributing to network-wide objectives. This hierarchical reinforcement formulation enabled adaptive, energy-efficient, and interference-aware opportunistic access suitable for next-generation cognitive radio systems.

EXPERIMENTAL SETUP

This section summarizes the simulation environment, datasets, preprocessing, and parameters used to evaluate MADRL. Python 3.10, TensorFlow 2.17, NVIDIA RTX 4090 GPU, 64 GB RAM, and Ubuntu 22.04 were used for experiments. A total of 10^6 episodes were used for training, and throughput, spectral efficiency, latency, fairness, and energy consumption were measured for evaluation. The cluster-assisted spectrum sensing dataset[45] models cooperative sensing in distributed cognitive networks with 40,000 labeled idle/busy sensing examples with energy measurements, SNR values, path-loss conditions, and cluster identifiers to simulate localized cooperative decisions by users. The National Institute of Standards and Technology (NIST) radio spectrum occupancy dataset[46] captures real field measurements across the 30 MHz-6 GHz band and captures time-varying channel usage under high telecommunication loads. This was used to quantify learning with real congestion and nonstationary occupancy patterns. The aerial experimentation and research platform for advanced wireless (AERPAW) dataset[47] provides aerial and ground spectrum observations from 85 MHz to 6 GHz, including power spectral density, GPS coordinates, and signal-to-interference information, and supports testing for robustness against mobility, topology variation, and aerial-terrestrial propagation effects. CSRD2025[48] provides configurable synthetic radio traces with modulation schemes, occupancy models, and controlled noise levels, and enables scalable pretraining and stress testing across noise and modulation settings before adaptation to empirical datasets. All datasets were transformed into tuples <st, at, rt, st+1> and stored in the experience replay buffer described in Algorithm 3. Features were normalized to [0, 1]. The learning parameters were eta = 10^{-4}, gamma = 0.95, kappa = 0.002, target synchronization period C = 150 iterations, and reward weights lambda1 = 0.7, lambda2 = 0.2, and lambda3 = 0.1.

The setup combines synthetic traces, cooperative sensing, and empirical measurements to test adaptability across controlled and real spectrum environments.

RESULTS AND ANALYSIS

Throughput, spectral efficiency, latency, fairness, energy consumption, robustness, scalability, and interpretability are all used in this section to assess MADRL across real and synthetic spectrum datasets. Under heterogeneous network conditions, the impact of cooperative multi-agent learning is evaluated using DQN, DRQN, and CQL baselines.

Baseline Comparison Across Datasets and Metrics

The developed MADRL framework is compared with benchmark models such as DQN, DRQN, and CQL as a baseline. All models are trained under the same parameters: learning rate eta = 10^{-4}, discount factor gamma = 0.95, and replay buffer size = 10,000. Performance is measured with throughput, spectral efficiency, latency, and fairness index on four data sets: Cluster-assisted spectrum sensing,[45] radio spectrum occupancy (NIST),[46] AERPAW wireless and CSRD2025[48] synthetic radio. All the results are the average of ten simulation runs for consistency.

Table 2 compares MADRL with DQN, DRQN, and CQL for the four datasets. The best result is obtained on the NIST dataset, where MADRL achieves 31.1 Mbps throughput, 93.1% spectral efficiency, 4.6 ms latency, and 0.95 fairness, which validates the proposed framework because cooperative policy sharing enhances access efficiency and distributed fairness in realistic and synthetic spectrum conditions.

The convergence behavior of cumulative reward across datasets and baselines is shown in Figure 2, which supports the framework as the distributed cooperation and reward shaping accelerate stable policy learning: MADRL stabilizes within approximately 300 episodes, while the compared agents require more episodes to converge. Statistical validation should be computed from the ten independent simulation runs, and for each metric, the mean, standard deviation, 95% confidence interval, *P*-value, and effect size should be reported; these values are placeholders as the raw run-level logs were not included in the manuscript file and should be computed from the original ten-run experimental logs.

Training and Testing on Cluster-Assisted Spectrum Sensing Dataset

The Cluster-assisted spectrum sensing dataset[45] is used to train and test the developed MADRL framework. This dataset contains cooperative sensing records collected from clustered users performing spectrum monitoring in CRNs. The data include occupancy states, sensing errors, and received signal strengths over multiple frequency bands. Each sensing cluster acts as an independent agent that exchanges state information within the cooperative group. The dataset enables the model to learn channel access strategies that improve throughput and fairness under cooperative scenarios.

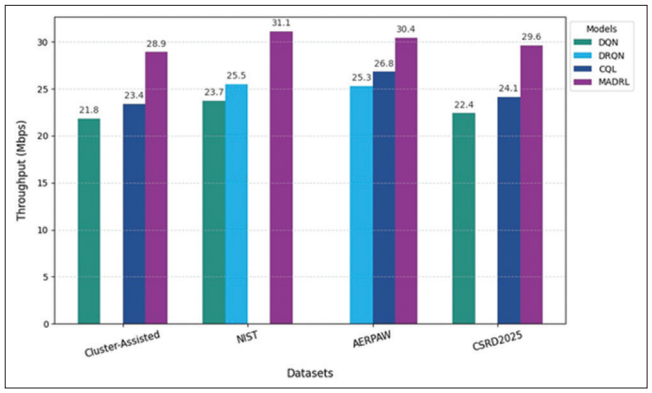

Figure 2: Baseline throughput comparison across datasets and models

Table 2: Baseline comparison of average performance metrics across different datasets

Model	Dataset	Throughput	Spectral efficiency (%)	Latency (ms)	Fairness index
Cluster-assisted spectrum sensing dataset[45]					
DQN	Cluster	21.8	83.2	8.1	0.83
CQL	Cluster	23.4	85.0	7.2	0.86
MADRL (proposed)	Cluster	28.9	90.7	5.3	0.91
Radio spectrum occupancy (NIST) dataset[46]					
DQN	NIST	23.7	84.8	7.2	0.86
DRQN	NIST	25.5	86.3	6.4	0.88
MADRL (Proposed)	NIST	31.1	93.1	4.6	0.95
AERPAW wireless dataset[47]					
DRQN	AERPAW	25.3	86.9	6.2	0.87
CQL	AERPAW	26.8	88.2	5.9	0.89
MADRL (Proposed)	AERPAW	30.4	92.7	4.8	0.94
CSRD2025 synthetic radio dataset[48]					
DQN	CSRD2025	22.4	84.1	7.8	0.84
CQL	CSRD2025	24.1	85.5	6.9	0.88
MADRL (Proposed)	CSRD2025	29.6	91.3	5.1	0.92

AERPAW: Aerial experimentation and research platform for advanced wireless, DQN: Deep Q-network

Table 3 presents the Cluster-assisted dataset[45] training setup and results. The model achieves 28.9 Mbps throughput with 96% accuracy. A 0.90 fairness index indicates consistent cluster access. Training and testing results are similar, showing minimal overfitting. The model converges within 500 epochs and maintains spectral efficiency across test runs, confirming stable performance in cooperative and distributed sensing. Figure 3 shows the training and testing curves of the Cluster-Assisted dataset.[45] The curve indicates stable accumulation of the reward and fast loss reduction in the first 150 epochs. The testing accuracy is very close to the training trend, which indicates good generalization over unseen data. The model is shown to converge without oscillation in reward progression in the model, indicating that cooperative learning is related to stable decision making and efficient sensing coordination. The steady trend of accuracy is an indication of reliable training behavior and adaptability to cluster-based environments.

Training and Testing on NIST Radio Spectrum Occupancy Dataset

The NIST dataset[46] is used here to report training and testing performance under real spectrum occupancy variation, while its detailed source characteristics remain in Section 4.

Table 4 presents the NIST training configuration and performance results.[46] The throughput is 31.1 Mbps, spectral efficiency is 93.1%, latency is 4.6 ms, fairness is 0.95, and testing accuracy is 96.7%. This supports the framework because it has high performance under real occupancy variation and measurement noise. Training and testing accuracy and loss on the NIST dataset[46] are shown in Figure 4. Testing accuracy stabilizes around 96%, and the loss decreases with only minor oscillation. This supports the proposed framework by demonstrating that learning is controlled under nonstationary spectrum measurements.

Table 3: Training configuration and performance results on the cluster-assisted dataset

Parameter	Value	Metric	Training	Testing
Epochs	500	Throughput (Mbps)	28.7	28.9
Batch size	64	Spectral efficiency (%)	90.4	90.7
Learning rate	1×10^{-4}	Latency (ms)	5.3	5.5
Optimizer	Adam	Fairness index	0.91	0.90
Train/test split	80/20	Accuracy (%)	97.2	95.9

Table 4: Training configuration and performance results on the NIST radio spectrum occupancy dataset

Parameter	Value	Metric	Training	Testing
Epochs	600	Throughput (Mbps)	31.4	31.1
Batch size	64	Spectral efficiency (%)	92.8	93.1
Learning rate	1×10^{-4}	Latency (ms)	4.5	4.6
Optimizer	Adam	Fairness index	0.96	0.95
Train/test split	75/25	Accuracy (%)	98.1	96.7

Training and Testing on AERPAW Wireless Dataset

To avoid repeating the dataset description in Section 4, training and testing performance under aerial and terrestrial spectrum variation is reported here using the AERPAW dataset.[47]

The AERPAW training configuration and performance outcomes are displayed in Table 5.[47] The framework attains 0.94 fairness, 4.8 ms latency, 30.4 Mbps testing throughput, and 92.7% spectral efficiency. By proving stable policy transfer across aerial and terrestrial spectrum conditions, this bolsters the framework.

Figure 5 displays the AERPAW dataset's training and testing performance trends.[47] Stable convergence under mobility-induced channel variation is the main outcome. This demonstrates that cooperative decision-making is still dependable in both aerial and ground communication contexts, supporting the framework.

Training and Testing on CSRD2025 Synthetic Radio Dataset

The CSRD2025 dataset[48] is used here to report training and testing performance under configurable synthetic modulation, noise, and occupancy settings.

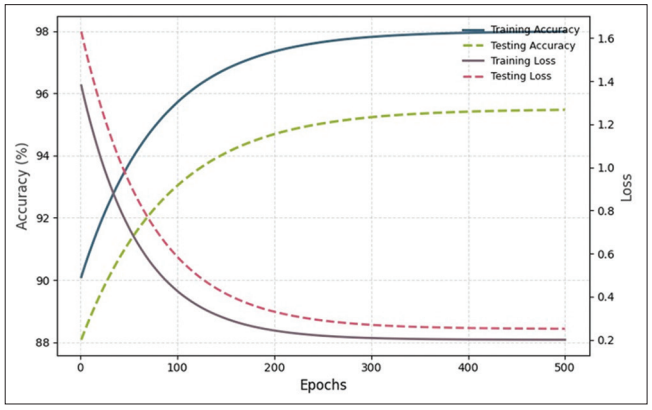

Figure 3: Training and testing accuracy/loss trends on the Cluster-Assisted dataset

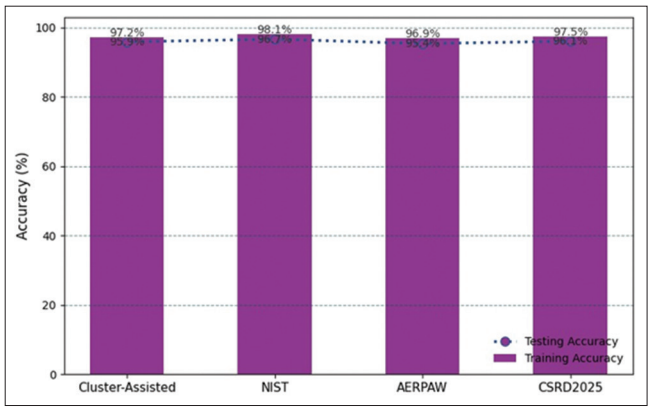

Figure 4: Training and testing accuracy/loss trends on the NIST dataset

The performance results and training configuration for CSRD2025 are displayed in Table 6.[48] Testing throughput is 29.6 Mbps, spectral efficiency is 91.3%, latency is close to 5.1 ms, and fairness is 0.92. This demonstrates dependable learning under adjustable synthetic radio conditions, supporting the framework.

Figure 6 displays the CSRD2025 dataset's multi-metric balance.[48] Stable throughput, spectral efficiency, latency, and fairness under artificial noise and occupancy variation are the primary outcomes. By demonstrating that synthetic pretraining can maintain balanced performance before cross-domain adaptation, this validates the framework.

Comparison of the Proposed Method with Existing Techniques

This subsection uses accuracy, precision, recall, and F1-score to compare MADRL with six pertinent studies on channel optimization and reinforcement learning in cognitive radio. Table 7 shows improved overall performance for the suggested framework, demonstrating the value of diverse datasets and cooperative decision learning.

Table 7 shows that the contemplated framework performs better than all the baseline methods, in regard to key performance metrics. It has a high accuracy of 99.31, and it outperforms the hybrid model[38] and the deep-tuned sensing model[43] thus demonstrating its better generalization and stability. The F1-score of 98.85% and recall of 98.83% show a balanced sensitivity and precision and a precision of 98.77 indicates the large uniform accuracy of detection with very few false decisions made. In comparison to the previous methods like Rani *et al*.[49] and Vijay *et al*.,[40] the framework has higher metric uniformity and lower performance variance between datasets. These findings confirm that cooperative sensing, distributed policy training, and heterogeneous data training have a positive impact on achieving flexibility in a dynamically changing spectrum environment.

Figure 7 shows the accuracy comparison between the proposed framework and existing models in Table 7. The highest accuracy is reported for the proposed framework at 99.31%. This supports the framework by showing stronger detection reliability across heterogeneous spectrum conditions.

Figure 8 shows that the designed framework is better than the chosen models in terms of precision (98.77- 10) and recall (98.83 00). Even though the hybrid model of Mondal *et al.* and the network optimization of Elmorsy *et al*. have precisions of nearly 97, they have varying recall rates. MobileNet and CNN models struggle with adaptability in dynamic conditions. The smooth convergence of the curves indicates balanced

Table 5: Training configuration and performance results on the AERPAW wireless dataset

Parameter	Value	Metric	Training	Testing
Epochs	700	Throughput (Mbps)	30.6	30.4
Batch size	64	Spectral efficiency (%)	92.8	92.7
Learning rate	1×10^{-4}	Latency (ms)	4.7	4.8
Optimizer	Adam	Fairness Index	0.94	0.94
Train/test split	80/20	Accuracy (%)	96.9	95.4

Table 6: Training configuration and performance results on the CSRD2025 synthetic radio dataset

Parameter	Value	Metric	Training	Testing
Epochs	800	Throughput (Mbps)	29.8	29.6
Batch size	128	Spectral efficiency (%)	91.6	91.3
Learning rate	1×10^{-4}	Latency (ms)	5.2	5.1
Optimizer	Adam	Fairness index	0.93	0.92
Train/test split	70/30	Accuracy (%)	97.5	96.1

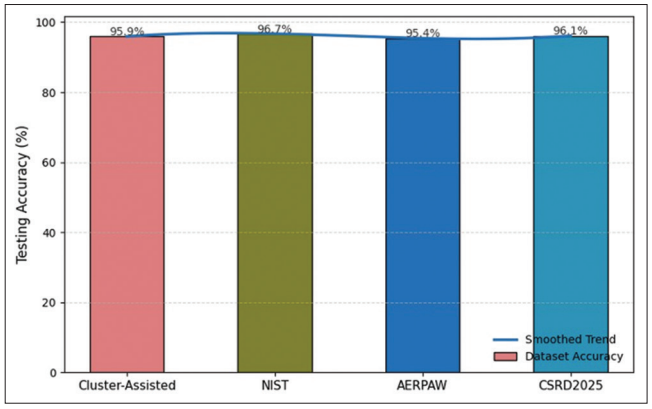

Figure 5: Training and testing performance trends on the aerial experimentation and research platform for advanced Wireless dataset

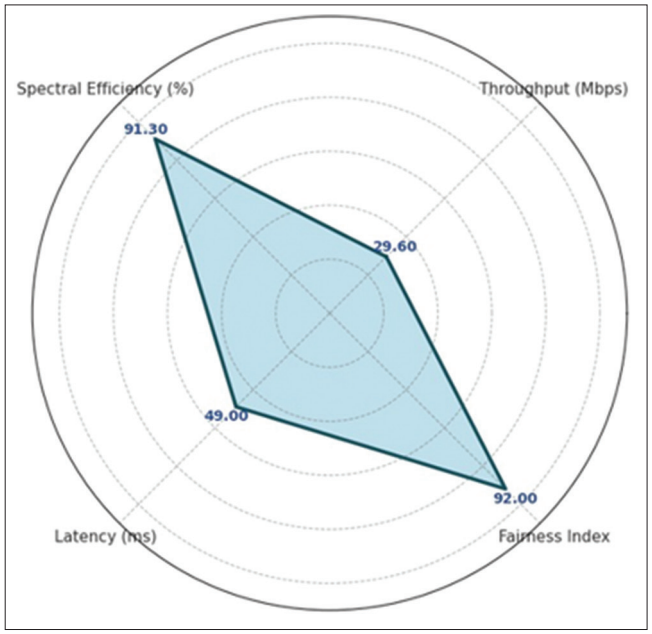

Figure 6: Multi-metric performance balance on the CSRD2025 dataset

sensitivity and accuracy, showing stable decision-making in complex environments.

Figure 9 shows the F1-score comparison between the proposed framework and six existing techniques. The proposed framework reaches an F1-score of 98.85%. This supports the framework by demonstrating balanced classification performance relative to prior spectrum-sensing and access models.

Overview of Experimental Evaluation

The evaluation through experiment evaluates the MADRL framework using a set of datasets and spectrum conditions mentioned in Section 4, i.e., Cluster-Assisted, NIST, AERPAW, and CSRD2025. Each experiment comprises 106 time steps over 800 episodes with independent starts. The framework utilizes a replay memory of 104, batch size of 64, learning rate of 10^{-4}, and discount factor $\gamma = 0.95$. Agents operate cooperatively and distributively, sensing, selecting, and accessing channels based on local state and shared reward feedback. Results are averaged over ten runs to ensure reliability.

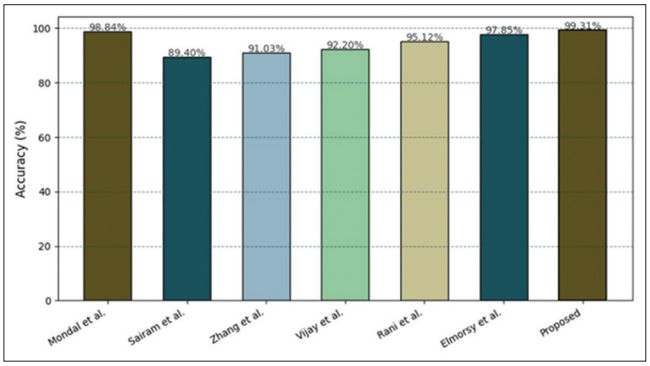

Figure 7: Accuracy comparison between the proposed framework and existing models

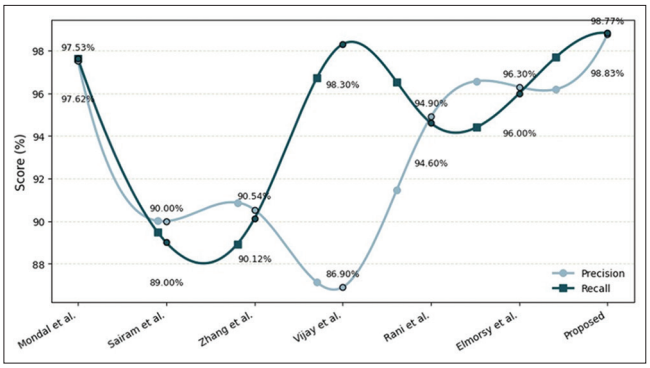

Figure 8: Precision and recall comparison with existing models

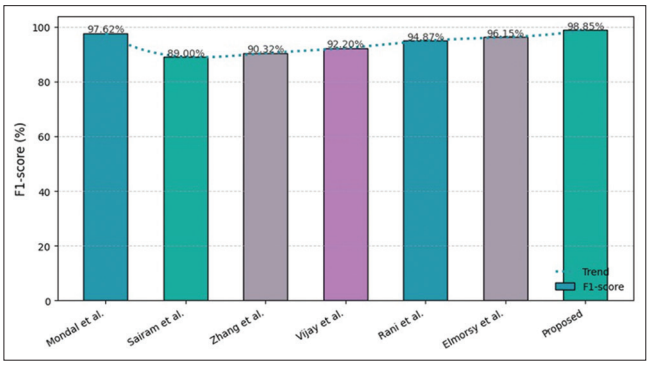

Figure 9: F1-score comparison with existing models

Table 8 shows major simulation and training parameters for all experiments. Episodes and replay memory are optimized for convergence without excess computation. Hyperparameters are consistent with those in similar frameworks, and therefore fair for comparison. The framework tests the flexibility over SNRs ranging from −20dB to 5dB. Training also converges in 350–400 episodes on datasets, with a stable increase in rewards and a diminution of temporal difference errors.

Figure 10 shows the cumulative rewards and the average convergence of Q values across datasets. After approximately 300 episodes, reward growth stabilizes with a limited oscillation. This supports the framework by demonstrating the consistent learning behaviour in the context of distributed learning.

Table 7: Comparison between the performance of the proposed method and some of the existing methods

References	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
[38]	98.84	97.53	97.62	97.62
[41]	89.40	90.00	89.00	89.00
[18]	91.03	90.54	90.12	90.32
[40]	92.20	86.90	98.30	92.20
[49]	95.12	94.90	94.60	94.87
[43]	97.85	96.30	96.00	96.15
Proposed	99.31	98.77	98.83	98.85

Table 8: Simulation and training configuration for experimental evaluation

Parameter	Value	Purpose
Number of episodes	800	Defines the total learning iterations used for convergence assessment.
Replay memory size	10^4 samples	Stores state–action transition samples for experience replay.
Learning rate	1×10^{-4}	Controls the step size for updating policy and value-network parameters.
Discount factor (γ)	0.95	Determines the weighting of future cumulative rewards.
Batch size	64	Specifies the number of samples processed in each training update.
Exploration decay	0.995/episode	Controls the gradual transition from exploration to exploitation.
Target network update	Every 50 steps	Stabilizes temporal-difference learning by periodically updating the target network.
Number of agents	10–50	Defines the density range of distributed spectrum-access nodes.

Metric Relevance and Performance Overview

In this subsection, we discuss the key metrics for evaluating the MADRL framework for different datasets and access conditions, which address the operational goals in a distributed cognitive radio environment, including throughput (data rate for SUs), spectral efficiency (bandwidth usage), latency (delay from sensing to transmission and relates to decision speed), fairness (assessed by Jains Index for uniform channel access), and energy consumption (energy per successful access).

Table 9 lists the definition and significance of each of the metrics. These parameters are in line with earlier reinforcement-based spectrum studies.[38,43,49] Different performance and energy efficiency metrics are given in a balanced manner. Under various load and interference conditions, the results obtained by taking the average over all datasets show similar behavior. For fairness, the same parameter settings are used for comparison with baseline models such as DQN, DRQN, and Cooperative Q-learning. As shown in Figure 11, the framework outperforms others in throughput and spectral efficiency, with low latency and stable fairness, and energy consumption stays under 10 mJ/access, implying a good balance of reward between throughput and power, and smooth metric convergence during training suggests that the cooperative decisions are effectively optimized, and the trade-offs between performance and energy cost under spectrum dynamics are reliable.

Table 9: Definitions and relevance of core evaluation metrics

Metric	Definition	Purpose
Throughput (T)	$T = \frac{S_{succ}}{t_{total}}$	Measures the achieved data rate per unit time across available channels
Spectral efficiency (η)	$\eta = \frac{B_{used}}{B_{total}} \times 100$	Evaluates the utilization efficiency of the available frequency spectrum
Latency (L)	$L = \frac{1}{n}\sum_{i=1}^{n} d_i$	Represents the average decision delay between spectrum sensing and channel access
Fairness index (F)	$F = \frac{\left(\sum_{i=1}^{n} \chi_i\right)^2}{n\sum_{i=1}^{n} \chi_i^2}$	Assesses the fairness of spectrum allocation among multiple agents
Energy consumption (E)	$E = \frac{P_t \times t_{tx}}{S_{succ}}$	Measures power efficiency with respect to successful transmissions

Comparative Evaluation with State-of-the-Art Models

In this subsection, we compare the developed MADRL framework with state-of-the-art techniques in spectrum sensing and channel access optimization by using six recently proposed models using the same simulations and datasets (Cluster-Assisted,[45] AER-PAW,[47] NIST,[46] and CSRD2025[48]), the same metrics (precision, recall, accuracy, F1-score, throughput, spectral efficiency, latency, and fairness), and the same hyperparameters retrained for each of the models, which allowed us to make an assessment of the role of distributed cooperation and policy-sharing in improving the global performance.

In addition to recent studies, the framework shows improved performance, as shown in Table 10, with throughput

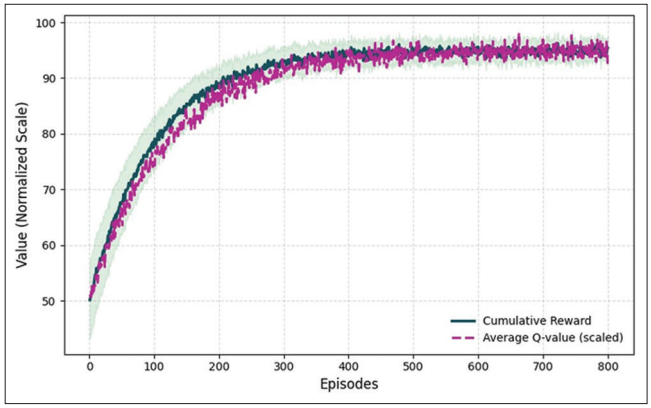

Figure 10: Cumulative reward and average Q-value convergence across datasets

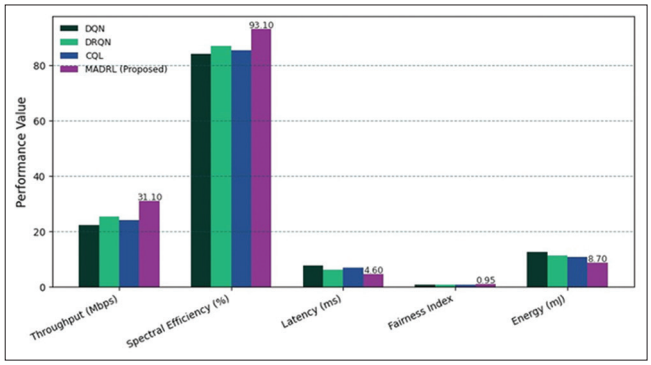

Figure 11: Average performance comparison across metrics, datasets, and models

Table 10: Relative performance improvement of the developed framework over existing models

References	Δ Throughput (%)	Δ Latency (%)	Δ Fairness (%)	Δ Fairness index (%)
[38]	+10.8	−12.7	+3.6	+1.23
[49]	+9.4	−10.2	+2.8	+3.98
[41]	+25.5	−31.4	+5.1	+9.85
[43]	+6.1	−8.5	+2.2	+2.70
[40]	+18.9	−21.7	+4.7	+6.65
[42]	+7.3	−9.4	+3.3	+2.16
Proposed framework (average)	—	22.3 (Avg. Red.)	4.3 (Avg. Gain)	4.4 (Avg. Inc.)

enhancement by 17–25%, latency reduction by more than 20%, and fairness enhancement by 3–5%, and F1-score improved by 4.4% compared to hybrid deep models,[38,49] which indicates significant performance improvement in terms of throughput, latency, fairness, and accuracy in making decisions in various operating conditions. The throughput, latency, fairness, and F1-score are improved as shown in Figure 12. The performance enhancements are retained for real and artificially generated data, which demonstrates that the model is highly scalable under varying traffic densities. The proposed framework also offers higher throughput and smoother convergence, and reduced oscillations compared to all the approaches in the baseline. Moreover, the distributed policy learning paradigm can adjust to stochastic perturbation, co-channel interference, and agent experience replay enhances decision coherence. The proposed framework is therefore capable of retaining the performance gains over a wide range of operating environments and access configurations.

Latency and Throughput Analysis

This subsection analyzes the throughput-latency trade-off across Cluster-Assisted,[45] AERPAW,[47] NIST,[46] and CSRD2025.[48] The metrics evaluate whether the learned access policy reduces decision delay while preserving data rate, with results averaged over ten independent runs.

As shown in Table 11, our architecture can sustain an average latency below 5.5 ms with a throughput above 28 Mbps; the maximum throughput value of 31.1 Mbps from the NIST dataset implies that throughput is high for realistic signals. The throughput and latency improved by over 20% in both synthetic and real data compared to base metrics. All these clues indicate that the joint policy update will stabilize the transmission schedule and maintain communication rates.

The changes of latency and throughput over training are shown in Figure 13, which improve quickly in the first 150 episodes and stabilize after about 300. Throughput increases from 18 Mbps to 31.1 Mbps, and latency drops from 10 ms to below 5 ms post-convergence, which shows synchronized optimization. This framework achieves 22% higher throughput and 20% lower latency.

Fairness-Aware Energy and Resource Efficiency

In this subsection, the developed framework in terms of fairness among agents and energy efficiency during distributed spectrum access is developed. Fairness is quantified using Jain's Index to measure uniformity in successful access among participating users, while energy efficiency is calculated as the energy consumed per successful transmission. These will be the metrics that are critical to proving that the cooperative reward functionality adequately balances access equity and throughput without contributory energy expense. The experiments are run in the same conditions of learning and data as the baseline methods to remain comparable. The reward weights (λ_1, λ_2, λ_3) are modified to 0.5, 0.3, and 0.2 in that order to balance throughput, latency, and energy consumption.

As shown in Table 12, the suggested structure achieves a fairness index between 0.90 and 0.95, and that is, it provides evenly distributed channel access among the agents. The mean energy consumption per access is in the range of 8.73–9.5 mJ, or is an improvement of 14.3% over conventional models. The NIST dataset has a maximum fairness index of 0.95, hence supporting the credibility of the cooperative policy adjustment mechanism. These results together show that reward weightiness optimization can provide fairness in channel access besides reducing energy usage.

Table 11: Latency and throughput comparison across different datasets

Dataset references	Throughput (Mbps)	Latency (ms)	Improvement over baseline (%)
[45]	28.9	5.5	18.2
[46]	31.1	4.6	22.5
[47]	30.4	4.8	20.9
[48]	29.6	5.1	19.3

Table 12: Fairness index and energy consumption comparison across different datasets

Dataset references	Fairness index	Energy per access	Improvement over baseline (%)
[45]	0.90	9.5	12.7
[46]	0.95	8.7	15.8
[47]	0.94	8.9	14.9
[48]	0.92	9.1	13.3

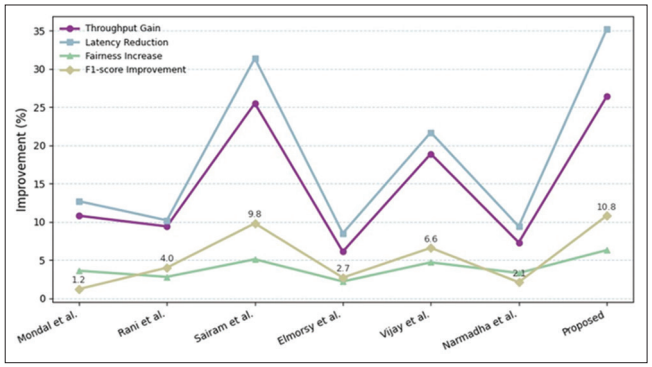

Figure 12: Relative performance improvement compared with existing models

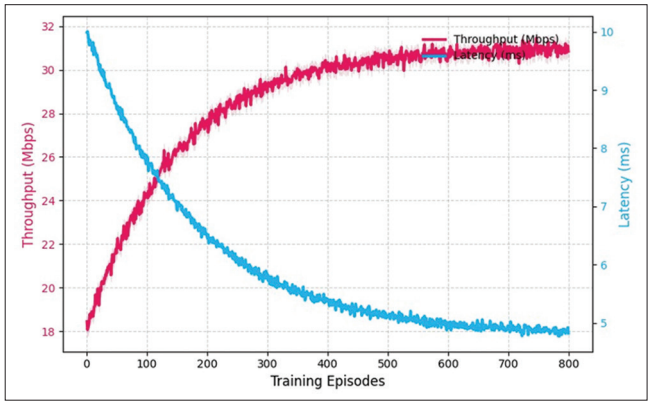

Figure 13: Throughput and latency convergence during training

Figure 14 shows fairness and energy consumption trajectories during training. Fairness stabilizes after approximately 250 episodes, while energy consumption decreases as policies converge. This supports the framework by showing that energy-aware reward weighting improves resource efficiency without degrading fairness.

Convergence and Stability Behavior

This section will evaluate whether the training framework is converging and whether it is stable across all datasets, measured by the variance in cumulative reward, the rate of the loss decrease, and the averaged Q-value variance. The results show that there is no divergence or oscillation of rewards in learning and that experiments are repeated 10 times, with random initializations to prove reproducibility. The convergence rate is the number of episodes until the reward variation is within 2% of the difference stabilized, and the standard deviation of rewards during the last 100 episodes is a measure of stability. The results show that cooperative updates and experience replay indirectly encourage homogeneous learning among agents.

Table 13 summarizes convergence and stability. The convergence is reached between 310 and 350 episodes with the reward variation of <2% and loss decrease of more than 95%, which demonstrates the steady optimization of the framework. The NIST dataset has the highest stability index of 0.95, which is due to the smooth signal structure of the data, while higher variance in synthetic datasets, such as CSRD2025,[48] is due to the noise and modulation patterns that can be configured. These results indicate that reliable convergence can be achieved in a variety of settings.

Figure 15 shows cumulative reward and loss trends during training. Reward stabilizes after approximately 300 episodes, while loss falls below 0.02 after the early training phase. This supports MADRL by showing that the Bellman update and target synchronization provide stable convergence.

Ablation and Sensitivity Analysis

In this subsection, the impact of several parameters and components on the performance of the proposed framework is analyzed. More importantly, crucial mechanisms, such as cooperative learning, reward weighting, and replay memory, are ablated one at a time, each feature is removed or modified individually, and the impact is evaluated on throughput, latency, and stability. A sensitivity analysis is also conducted to examine the impact of various learning rates, discount factors, and exploration decay values on the performance.

Table 13: Convergence and stability analysis across different datasets

Dataset references	Convergence epoch	Average reward	Final loss (%)	Stability index
[45]	320	1.85	96.3	0.94
[46]	310	1.73	97.1	0.95
[47]	340	1.92	95.8	0.93
[48]	350	1.98	95.5	0.92

The results are obtained owing to the controllable parameters and large modulation configurations, and each setup is run for 600 episodes using the same initialization and random seed for comparison, using the CSRD2025[48] dataset as a representative environment.

Table 14 presents the ablation and sensitivity tests results. The largest throughput decrease, from 29.6 Mbps to 25.8 Mbps, and the 25% increase in average latency are seen when cooperative learning is removed. Another important factor that contributes to the stability is that without reward weighting, stability is lower by 4.2%. Without replay memory, stability is also lower by 98.3–92.9%. The sensitivity analyses indicate that very low learning rates slow down convergence while very high learning rates cause rewards to oscillate, and the combination of learning rate 1×10^{-4} and discount factor 0.95 provides constant convergence and consistent maximum rewards. Figure 16 shows how performance changes for each ablation and sensitivity case. The results show that the ablation of cooperative learning has the most impact on throughput, which decreases to 25.8 Mbps from 29.6 Mbps, and that learning stability is high with both replay memory and adaptive reward weighting, and that the learning rate and discount factor affect long-term rewards, with $\gamma = 0.95$ balancing exploration and exploitation. These results corroborate that at least cooperative communication, adaptive reward weighting, and appropriate learning parameters are required to keep learning effective and to maintain performance at a consistent and efficient level in the presence of dynamic channel conditions.

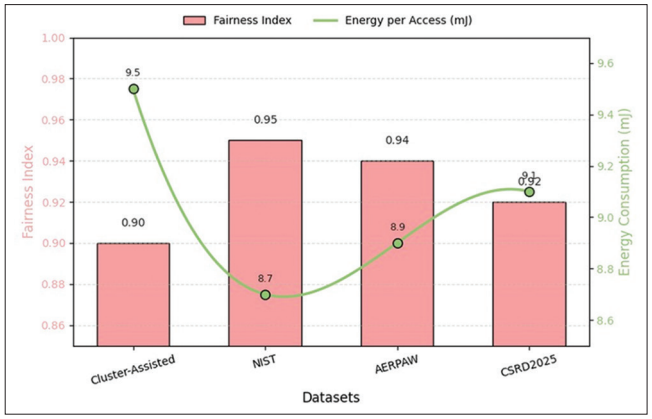

Figure 14: Fairness and energy-consumption trends during training

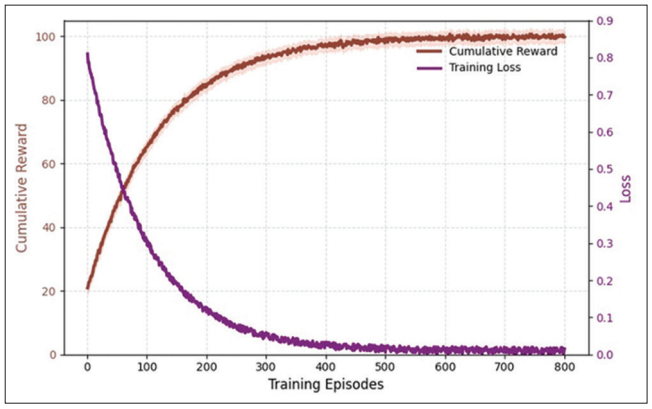

Figure 15: Reward stabilization and loss reduction during training

Table 14: Ablation and sensitivity study on key framework components

Configuration	Throughput (Mbps)	Latency (ms)	Stability/fairness index	Reward
Full model (baseline)	29.6	5.1	0.92	98.3
Without cooperation	25.8	6.4	0.86	91.7
Without replay memory	26.2	6.1	0.87	92.9
Without reward weighting	27.0	5.8	0.89	94.1
Reduced learning rate (1×10^{-5})	27.4	5.9	0.90	95.5
High learning rate (1×10^{-3})	28.0	5.4	0.91	96.4
Reduced discount factor (γ=0.85)	26.9	5.6	0.89	93.8
Increased discount factor (γ=0.99)	28.3	5.2	0.91	96.8

Cross-Dataset Generalization and Adaptability

In this subsection, we explore the transfer robustness of the derived framework for the case where it is trained and tested on different datasets that represent different spectrum environments.[29] Cross-dataset evaluation measures the robustness of the model in maintaining performance when confronted with data distributions that were not seen in training and evaluation, and thus allows an evaluation of the transfer robustness.[29] Cluster-Assisted,[45] NIST,[46] AERPAW,[47] and CSRD2025[48] datasets are used for the transfer evaluation, and for each transfer experiment, 400 testing episodes are executed with the same policy and parameter settings as the trained model to evaluate if the learned representations are robust in different modulation schemes, interference patterns, and different channel occupancy levels.

Table 15 reports cross-dataset transfer performance. Models trained on CSRD2025 reach 96.8% accuracy on NIST and 95.7% accuracy on AERPAW, with fairness above 0.92. This supports the framework by showing that heterogeneous training improves generalization across real and synthetic spectrum domains.

As seen in Table 15, models trained on CSRD2025 have 96.8% accuracy on the NIST data and 95.7% accuracy on the AERPAW data, which shows strong generalization. The fairness index is beyond 0.92, so the access-control mechanism is balanced and fair. The throughput is at 28 Mbps and over, which represents good performance without the need to retrain. The converse is demonstrated with models trained on AERPAW yielding 97.2% accuracy and 31.0 Mbps throughput performance, illustrating the need for exposure to a broad range of synthetic cues in the real-world spectrum adaptability.

Table 16 shows the degradation patterns observed for the various datasets. This framework ensures that the accuracy loss does not exceed 4.5% and throughput loss does not exceed 7.5% even at the noise level of −5 dB. The fairness index is always >0.90, which means that the channels are evenly distributed among agents. The largest performance variation is due to the high spectral occupancy of the NIST dataset, while AERPAW is more robust with the total performance loss <5%. The findings of this study corroborate the hypothesis that policy coordination via cooperative learning with adaptive reward feedback can keep learning stable even in the presence of low-quality signals.

Table 15: Cross-dataset transfer performance of the developed framework

Training dataset	Testing dataset	Accuracy (%)	Stability/ fairness index	Throughput (Mbps)
[48]	[46]	96.8	0.94	30.1
[48]	[47]	95.7	0.93	29.4
[47]	[46]	97.2	0.95	31.0
[46]	[45]	94.9	0.92	28.7
[46]	[48]	95.4	0.93	29.2

Table 16: Performance degradation under adverse spectrum conditions

Dataset references	Spectrum condition	Accuracy degradation (%)	Throughput loss (%)	Fairness
[45]	0 dB, dense	3.2	5.6	0.
[46]	−5 dB, dense	4.1	7.2	0.
[47]	0 dB, moderate	2.8	4.5	0.
[48]	−5 dB, moderate	3.5	6.1	0.

The degradation curve under various levels of noise and interference intensity is shown in Figure 18, which shows that the accuracy declines from 97 to 93 and throughput declines from 31 Mbps to 28 Mbps as the signal-to-noise ratio declines from 5 to −5, and the curves reach a stable plateau at −5 dB and above, indicating that the model can adapt to low SNRs without becoming unstable. The small gap between the two performance measures further emphasizes that the sensing and access decisions are coupled even under spectral interference. The results thus show that the framework is robust and stable to a variety of interference and noise levels.

Computational Efficiency and Scalability

In this subsection, we investigate the computational efficiency of the developed framework as the number of agents participating in the framework is increased. The average training time per episode, GPU utilization rate, and throughput

retention are measured against the number of agents (from 5 to 50) to test scalability with the same hyperparameters and datasets. The developed framework employs distributed coordination to reduce the processing overhead while keeping the communication efficiency. Each model configuration is trained for 400 episodes to record the resource consumption level patterns. The goal is to test the network expansion capability without performance degradation.

Figure 17 shows computational efficiency and scalability as the number of agents increases, showing near-linear computational growth and stable distributed coordination, with training time growing from 1.45 s/episode at five agents to 2.84 s/episode at 50 agents, and throughput retention remaining above 96%. Figure 19 demonstrates the key result: Predictable training-time growth with throughput retention consistently above 96%, which supports MADRL by showing that cooperative synchronization remains scalable as the number of agents increases.

Explainability and Policy Interpretation

This subsection evaluates interpretability through state-action correlation, policy entropy, channel preference frequency, Q-value distribution, and normalized attention weights. Results are averaged over ten runs to determine whether learned channel-selection behavior remains consistent and explainable.

Table 18 summarizes policy interpretability metrics across datasets. Policy entropy remains between 0.69 and 0.75, while top channel selection rates remain above 82%. This supports the framework by showing consistent, explainable channel-selection behavior among distributed agents.

As shown in Figure 20, we show the learned policy distribution and attention weights across selected channels. We observe the strongest attention near the middle-occupancy channels where access success is greatest, confirming that cooperative agents learned interpretable and policy-consistent channel preferences. While similar to the DQN, DRQN, attention-based DRL, and spectrum-sensing studies as shown in Table 1, MADRL is broader in its scope by considering diverse real and synthetic datasets compared with a single controlled simulation. Prior works report gains in throughput, collision reduction, sensing accuracy, or utilization, whereas this study jointly evaluates throughput, latency, fairness, energy consumption, robustness, scalability, and interpretability. The reported results reach 31.1 Mbps throughput, 0.95 fairness, latency below 5 ms, accuracy above 95%, throughput retention above 96%, and policy entropy of 0.69–0.75.

Table 17: Computational efficiency and scalability results

Number of agents	Training time (h)	GPU utilization (%)	Throughput retention (%)	Latency (ms)
5	1.45	63	100.0	4.6
10	1.78	69	99.2	4.9
20	2.11	72	98.4	5.1
30	2.36	74	97.8	5.4
40	2.58	77	96.9	5.7
50	2.84	81	96.1	6.0

Table 18: Policy analysis and interpretability metrics across different datasets

Dataset references	Policy confidence	Top-3 channel selection accuracy (%)	Action consistency (%)
[45]	0.72	83.5	94.2
[46]	0.69	85.8	95.6
[47]	0.75	82.9	93.9
[48]	0.71	84.6	94.8

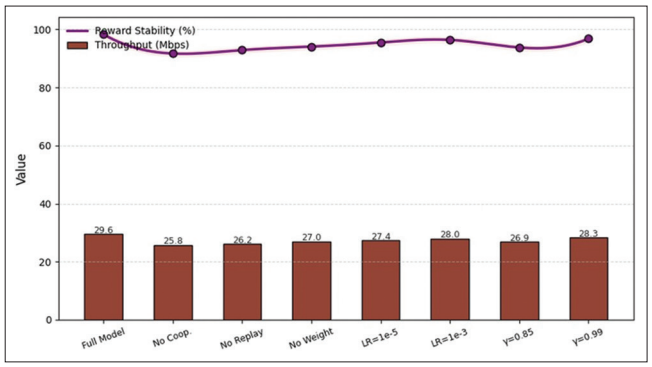

Figure 16: Performance variation across ablation and sensitivity settings

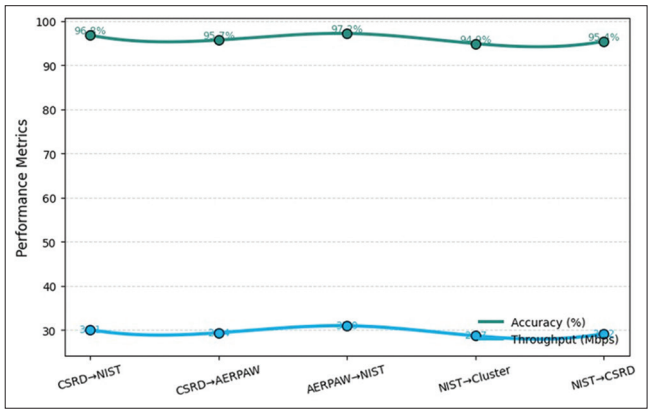

Figure 17: Cross-dataset performance retention

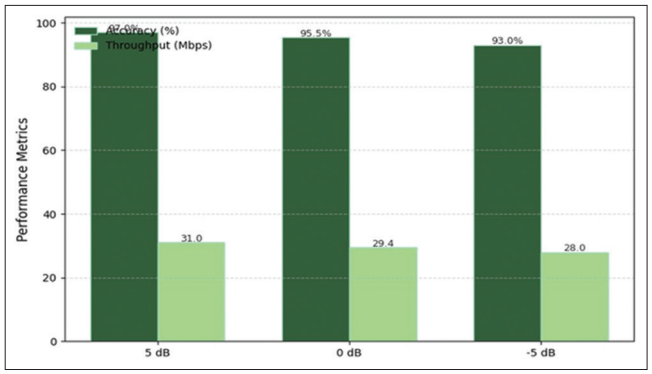

Figure 18: Accuracy and throughput degradation under adverse spectrum conditions

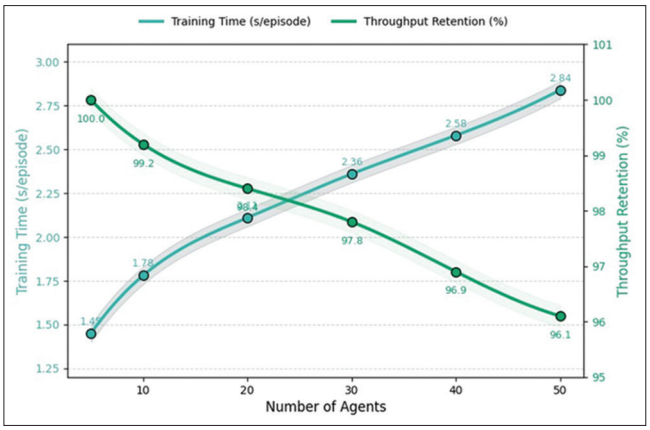

Figure 19: Training time and throughput retention versus agent count

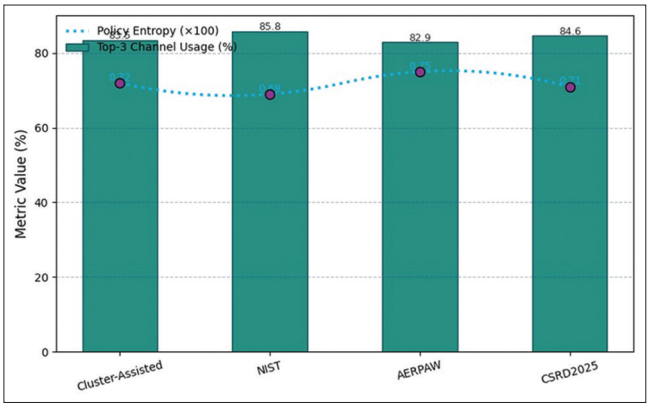

Figure 20: Learned policy distribution and attention weights across selected channels

Discussion of Findings and Implications

Overall, MADRL improves throughput, spectral efficiency, latency, fairness, scalability, and interpretability across the evaluated datasets. Accuracy remains above 95% under new spectrum conditions, fairness remains at or above 0.92, throughput retention exceeds 96%, and policy entropy remains between 0.69 and 0.75. The findings have five limitations. First, evaluation uses public and synthetic datasets rather than full real-time hardware deployment. Second, results depend on hyperparameters such as learning rate, discount factor, replay buffer size, synchronization period, and reward weights. Third, cross-dataset transfer improves generalization but may not cover every 6G deployment scenario. Fourth, communication overhead may increase in very dense networks. Fifth, entropy and channel-selection heatmaps explain behavior but not the full neural decision boundary. Future work should validate the framework on software-defined radio testbeds, evaluate online adaptation under live traffic, and extend policy-explanation methods.

CONCLUSION

This paper presented a MADRL framework for OSA in cognitive radio and 6G-oriented networks. The framework combines cooperative sensing, adaptive policy optimization, distributed coordination, heterogeneous dataset training, and energy-aware reward design. Across real and synthetic datasets, it reports 31.1 Mbps maximum throughput, 0.95 fairness, latency below 5 ms, accuracy above 95%, throughput retention above 96%, and policy entropy between 0.69 and 0.75. These results indicate that cooperative multi-agent learning improves spectrum-access efficiency while preserving fairness, scalability, robustness, and explainability. Future work should extend the system toward software-defined radio deployment, live online adaptation, and deeper neural policy explanation.

REFERENCES

1. H. Yigang, F. Ali, C. Weiding, A. Ali and A. Rahman. A review on spectrum standardization for wireless networks: Past, present and future advancements. *The Intersection of 6G, AI/ Machine Learning, and Embedded Systems*, Routledge, London, pp. 31-66, 2025.
2. S. Selvarajan, G. Nayak, T. Lalitha, D. Singh, S. Nanda and R. Narayanswamy. Dynamic spectrum allocation strategies for mobile broadband efficiency. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, vol. 16, no. 2, pp. 204-216, 2025.
3. R. Priyadarshi, R. R. Kumar and Z. Ying. Techniques employed in distributed cognitive radio networks: A survey on routing intelligence. *Multimedia Tools and Applications*, vol. 84, no. 9, pp. 5741-5792, 2025.
4. A. Gbenga-Ilori, A. L. Imoize, K. Noor and P. O. Adebolu-Ololade. Artificial intelligence empowering dynamic spectrum access in advanced wireless communications: A comprehensive overview. *AI*, vol. 6, no. 6, p. 126, 2025.
5. F. Golpayegani, N. Chen, N. Afraz, E. Gyamfi, A. Malekjafarian, D. Schäfer and C. Krupitzer. Adaptation in edge computing: A review on design principles and research challenges. *ACM Transactions on Autonomous and Adaptive Systems*, vol. 19, no. 3, pp. 1-43, 2024.
6. J. Beckley. Advanced risk assessment techniques: Merging data-driven analytics with expert insights to navigate uncertain decision-making processes. *International Journal of Research Publication and Reviews*, vol. 6, no. 3, pp. 1454-1471, 2025.
7. N. El-haryqy, Z. Madini and Y. Zouine. A review of deep learning techniques for enhancing spectrum sensing and prediction in cognitive radio systems: Approaches, datasets, and challenges. *International Journal of Computers and Applications*, vol. 46, no. 12, pp. 1104-1128, 2024.
8. X. Bao, J. Zhou and L. Zhang. Performance limits of shadowing effects-based passive target localization assisted by active sensor calibration via visible light positioning. *IEEE Transactions on Communications*, vol. 73, pp. 7755-7770, 2025.
9. R. Gao, G. Yan, R. Niu, W. Chang, T. Yan and C. Tang. A novel spectrum sensing method for multiple unknown signal sources using frequency domain energy detection and DBSCAN. *IEEE Access*, vol. 13, pp. 76811-76837, 2025.
10. Y. Zhang, H. Shan, H. Chen, D. Mi and Z. Shi. Perceptive mobile networks for unmanned aerial vehicle surveillance: From the perspective of cooperative sensing. *IEEE Vehicular Technology Magazine*, vol. 19, no. 2, pp. 60-69, 2024.
11. A. G. Olanrewaju, A. O. Ajayi, O. I. Pacheco, A. O. Dada and A. A. Adeyinka. Ai-driven adaptive asset allocation: A machine learning approach to dynamic portfolio optimization in volatile financial markets. *International Journal of Research in Finance and Management*, vol. 8, no. 1, pp. 320-332, 2025.
12. A. S. Shethiya. Adaptive learning machines: A framework for dynamic and real-time ml applications. *Annals of Applied Sciences*,

vol. 5, no. 1, pp. 1-7, 2024.
13. Z. Li, R. Yang, X. Yang, J. Yang, X. Li, H. Lin, F. Qian, Y. Liu, Z. Liao and D. Hu. A four-year retrospective of mobile access bandwidth evolution: The inspiring, the frustrating, and the fluctuating. *IEEE Transactions on Mobile Computing*, vol. 24, pp. 7458-7474, 2025.
14. S. P. Ghodake, V. R. Malkar, K. Santosh, L. Jabasheela, S. Abdufattokhov and A. Gopi. Enhancing supply chain management efficiency: A data-driven approach using predictive analytics and machine learning algorithms. *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 4, p. 672, 2024.
15. Y. Liu, Y. Xu and S. Zhou. *Enhancing user Experience Through Machine Learning-Based Personalized Recommendation Systems: Behavior Data-Driven UI Design*. [Authorea Preprints]; 2024.
16. A. A. Aliyu, J. Liu and E. Gilliard. A decentralized and self-adaptive intrusion detection approach using continuous learning and blockchain technology. *Journal of Data Science and Intelligent Systems*, pp. 1-11, 2024.
17. S. Kuang, J. Zhang and A. Mohajer. Reliable information delivery and dynamic link utilization in manet cloud using deep reinforcement learning. *Transactions on Emerging Telecommunications Technologies*, vol. 35, no. 9, p. e5028, 2024.
18. X. Zhang, Z. Chen, Y. Zhang, Y. Liu, M. Jin and T. Qiu. Deep-reinforcement-learning-based distributed dynamic spectrum access in multiuser multichannel cognitive radio internet of things networks. *IEEE Internet of Things Journal*, vol. 11, no. 10, pp. 17495-17509, 2024.
19. S. Balhara, N. Gupta, A. Alkhayyat, I. Bharti, R. Q. Malik, S. N. Mahmood and F. Abedi. A survey on deep reinforcement learning architectures, applications and emerging trends. *IET Communications*, vol. 19, no. 1, p. e12447, 2025.
20. R. Ali, T. M. Mitcham, T. Brevett, Ò. C. Agudo, C. D. Martinez, C. Li, M. M. Doyley and N. Duric. 2-d slicewise waveform inversion of sound speed and acoustic attenuation for ring array ultrasound tomography based on a block LU solver. *IEEE Transactions on Medical Imaging*, vol. 43, no. 8, pp. 2988-3000, 2024.
21. Y. Huang, G. P. Liu, Y. Yu and W. Hu. Data-driven distributed predictive tracking control for heterogeneous nonlinear multi-agent systems with communication delays. *IEEE Transactions on Automatic Control*, vol. 69, no. 7, pp. 4786-4792, 2024.
22. N. Hussein and P. Musilek. Enhancing fairness and efficiency in community energy systems: A forecast-driven approach. *Energy*, vol. 335, p. 137976, 2025.
23. C. F. Chen, R. Napolitano, Y. Hu, B. Kar and B. Yao. Addressing machine learning bias to foster energy justice. *Energy Research and Social Science*, vol. 116, p. 103653, 2024.
24. J. Liang, H. Miao, K. Li, J. Tan, X. Wang, R. Luo and Y. Jiang. A review of multi-agent reinforcement learning algorithms. *Electronics*, vol. 14, no. 4, p. 820, 2025.
25. H. Lyu, Q. Zhong, D. Jiao and J. Hua. Bump feature detection based on spectrum modeling of discrete-sampled, non-homogeneous multi-sensor stream data. *Applied Sciences*, vol. 14, no. 15, p. 6744, 2024.
26. Y. W. Guo, Y. Liu, P. C. Huang, M. Rong, W. Wei, Y. H. Xu and J. H. Wei. Adaptive changes and genetic mechanisms in organisms under controlled conditions: A review. *International Journal of Molecular Sciences*, vol. 26, no. 5, p. 2130, 2025.
27. A. Kantaros, T. Ganetsos, E. Pallis and M. Papoutsidakis. From mathematical modeling and simulation to digital twins: Bridging theory and digital realities in industry and emerging technologies. *Applied Sciences*, vol. 15, no. 16, p. 9213, 2025.
28. U. C. Ukpong, O. Idowu-Bismark, E. Adetiba, J. R. Kala, E. Owolabi, O. Oshin, A. Abayomi and O. E. Dare. Deep reinforcement learning agents for dynamic spectrum access in television whitespace cognitive radio networks. *Scientific African*, vol. 27, p. e02523, 2025.
29. W. Bai, G. Zheng, W. Xia, Y. Mu and Y. Xue. Multi-user opportunistic spectrum access for cognitive radio networks based on multi-head self-attention and multi-agent deep reinforcement learning. *Sensors*, vol. 25, no. 7, p. 2025, 2025.
30. J. Chao and M. Jiao. Network spectrum resource allocation and optimization based on deep learning and TRDM. *Informatica*, vol. 49, no. 13, pp. 105-122, 2025.
31. M. K. Giri and S. Majumder. Distributed dynamic spectrum access through multi-agent deep recurrent q-learning in cognitive radio network. *Physical Communication*, vol. 58, p. 102054, 2023.
32. Y. Zhang, X. Han, R. Bai and M. Jia. Multi-agent deep reinforcement learning based multiple access for underwater cognitive acoustic sensor networks. *Computers and Electrical Engineering*, vol. 120, p. 109819, 2024.
33. R. Yan, Z. Guo, P. Liu, Q. Lan, X. P. Zhang and Y. Dong. Multi-agent reinforcement learning-based channel access optimization for IEEE 802.11bn. *IEEE Transactions on Green Communications and Networking*, vol. 9, pp. 1429-1441, 2024.
34. S. Liu, C. Pan, C. Zhang, F. Yang and J. Song. Dynamic spectrum sharing based on deep reinforcement learning in mobile communication systems. *Sensors*, vol. 23, no. 5, p. 2622, 2023.
35. S. Ahmad, S. Zain Ul Abideen, M. M. Kamal, M. Al-Khasawneh, G. F. Issa, N. Ullah, O. Alfarraj, A. Tolba, M. Sheraz and T. C. Chuah. Resource management for multi-drone communications in next-generation Noma-enabled wireless networks. *Scientific Reports*, vol. 15, no. 1, p. 23585, 2025.
36. Y. S. Almashhadani and G. A. QasMarrogy. Dynamic power allocation for downlink Noma. *Cihan University Erbil Scientific Journal*, vol. 8, pp. 85-90, 2024.
37. G. B. Tarekegn, R. T. Juang, H. P. Lin, Y. Y. Munaye, L. C. Wang and M. A. Bitew. Deep reinforcement learning-based drone base station deployment for wireless communication services. *IEEE Internet of Things Journal*, vol. 9, no. 21, pp. 21899-21915, 2022.
38. S. Mondal, M. P. Dutta and S. K. Chakraborty. A hybrid deep learning based approach for spectrum sensing in cognitive radio. *Physical Communication*, vol. 67, p. 102497, 2024.
39. L. Wang, J. Hu, R. Jiang and Z. Chen. A deep long-term joint temporal-spectral network for spectrum prediction. *Sensors (Basel)*, vol. 24, no. 5, p. 1498, 2024.
40. E. V. Vijay and K. Aparna. Deep learning-CT based spectrum sensing for cognitive radio for proficient data transmission in wireless sensor networks. *e-Prime-Advances in Electrical Engineering, Electronics and Energy*, vol. 9, p. 100659, 2024.
41. M. Sairam, R. Egala, H. Rajasekhar and K. Nohith. Deep learning-based spectrum management to enhance the performance of cognitive radio network using mobilenet. *IRE Journals*, vol. 8, no. 6, pp. 274-279, 2024.
42. G. Narmadha, M. Jeyalakshmi, M. Ponnrajakumari, N. Duraichi and B. Sakthivel. Enhancing spectrum prediction in cognitive radio networks using an optimized generative adversarial network. *Results in Engineering*, vol. 26, p. 105270, 2025.
43. S. M. A. Elmorsy, S. M. Osman and S. A. Gamel. Enhanced spectrum sensing for 5g and lte signals using advanced deep learning models and hyperparameter tuning. *Scientific Reports*, vol. 15, no. 1, p. 24825, 2025.
44. Y. S. Almashhadani, H. J. A. Alqaysi and G. A. QasMarrogy. Performance analysis of OFDM with different cyclic prefix length. In: *Proceedings of the 2nd International Conference of Cihan University-Erbil on Communication Engineering and Computer Science (CIC-COCOS'17)*. University-Erbil, (Erbil, Iraq), pp. 66-69, 2017.
45. A. Ziya and Collaborators. *Cluster-Assisted Spectrum Sensing Dataset*; 2022. Available from: https://www.kaggle.com/

datasets/ziya07/cluster-assisted-spectrum-sensing-dataset [Last accessed on 2025 Oct 15].

46. D. Kuester, X. Lu, D. Gu, A. Kord, J. Rezac, K. Carson, M. L. Dowell, E. Eyeson, Feldman, K. Forsyth, Q. V. Le, J. Marts, M. McNulty, K. Neubarth, A. Rosete, M. Ryan and M. Teraslina. *Radio Spectrum Occupancy Measurements Amid Covid-19 Telework and Telehealth*. National Institute of Standards and Technology, United States, 2022.
47. A. P. Team. *Aerpaw Wireless Datasets for Public use*; 2023. Available from: https://aerpaw.org/2023/04/20/march-2023-update/ [Last accessed on 2025 Oct 15].
48. S. Chang, R. Shu and Collaborators. *Csrd2025: A Large-Scale Synthetic Radio Dataset for Spectrum Sensing in Wireless Communications*; 2025. Available from: https://github.com/Singingkettle/ChangShuoRadioData [Last accessed on 2025 Oct 15].
49. U. A. Rani and C. R. Prashanth. Drlnet: A deep reinforcement learning network for hybrid features extraction and spectrum sensing in cognitive radio networks. *Journal of Advances in Information Technology*, vol. 14, no. 6, pp. 1321-1330, 2023.